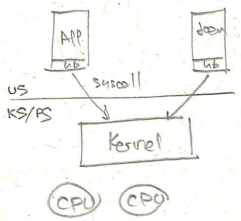
PART 1 – OPERATING SYSTEM

SIMONE GUGGIARI
ETH ZÜRICH 2016
LEGI # 14-932-867

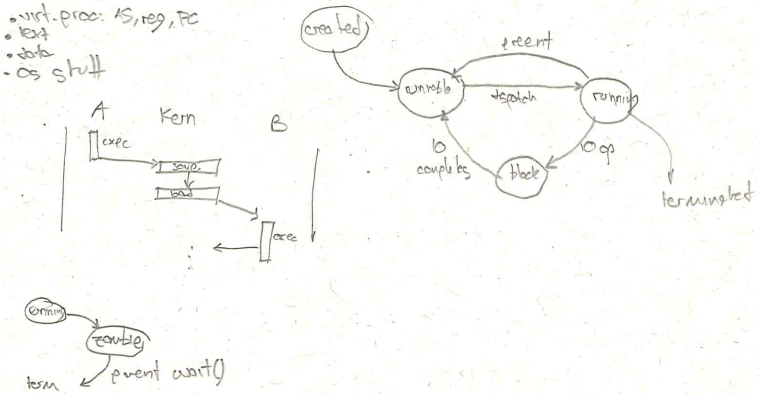
1 INTRODUCTION

- OS roles
  - referee
    - manages resource access
  - illusionist
    - presents as perfect virtual
  - glue
- Services
- Privileged mode
  - kernel / CPU
  - enter/exit conditions
- User mode
  - application / system library / daemon



2 PROCESSES

- Concept
- Lifecycle
  - state machine
- Switching
  - PCB
- Creation
  - requirements
    - fork(), exec() - copy - new
    - pid
    - -> see code
  - zombie
- Kernel threads
  - 1:1, M:1, M:N
- Kernel architecture
- System calls
  - steps 1-15
- User-space threads
  - multiplicity
    - pro/con

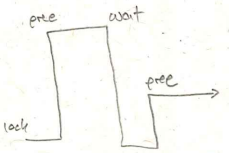


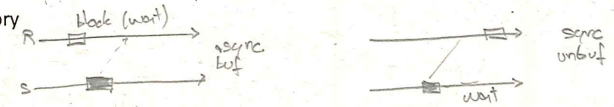
3 SCHEDULING

- Allocation
- problems:
  - complexity
  - frequency
  - policies to enforce
  - assumptions
- points in time
- Preemption
- Realtime systems
  - soft
  - hard
- Overhead
- BATCH oriented scheduling
  - properties/goals
  - FIFO
  - SJF / SRTF
    - & preemption = starvation
    - optimality
- INTERACTIVE loads scheduling
  - properties/goals
  - RR
  - Multi-level queues → consume quantum = lose priority
  - CFS (completely fair scheduler)
  - resource containers
- (Soft) real time / multimedia
  - properties/goals
  - RMS (rate monotonic scheduling) $U = \sum_i \frac{c_i}{p_i} \leq n(2^{1/n} - 1)$
  - formula
  - EDF
- Multiprocessor scheduling
  - 1. sequential programs
    - affinity scheduling
  - 2. parallel applications
    - co-scheduling
  - Little's Law

4 SYNCHRONIZATION

- Hardware support
  - interrupts disable
  - TAS
  - CAS
  - LL / SC
  - (competitive) spinning
- IPC (inter-process communication)
  - priority inversion
  - solutions
    - priority inheritance
    - priority ceiling
- IPC without shared memory
  - asynchronous
    - buffered
  - synchronous
    - unbuffered
  - pipes
    - -> see code
  - messaging system





unidir, half comm. chan between 2 proc

→ allows process communication (unidir) w/r
buffer → synchronization when full/empty

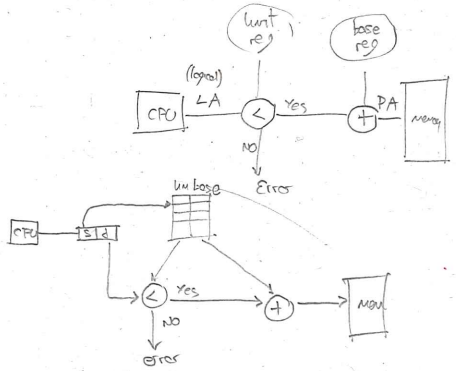
mk_mod()

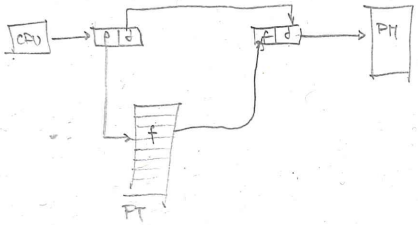
- requirements
- ports
- local RPC (remote procedure call)

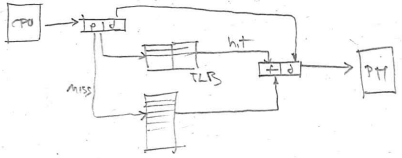
- Unix signals
  - SIG - HUP, INT, QUIT, KILL, SEGV, PIPE, ALRM, CHLD, STOP, CONT, SR1,2, TERM
    - shell
- Up-calls

5 MEMORY MANAGEMENT

- goals
- terminology
- MMU
- partitioned memory:
  - base/limit regs
  - contiguous allocation
- Segmentation
  - architecture
    - STBR / STLR
- Paging
  - terminology
  - frames / pages
  - P6
    - schema
    - translation
    - performance
    - cache PTE
    - (TLB)
  - x86-64
    - schema
    - translation
  - EAT (effective access time)
- Page protection
  - AVAIL, G, D, A, CD, WT, US, RW, P
- Page sharing
  - per-process protection
- Page table structures
  - hashed PT
  - inverted PT
- TLB shootdown
  - hardware coherence
  - virtual caches







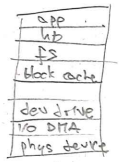
6 DEMAND PAGING

- virtual memory
  - advantages
- COW (copy on write)
  - vfork()
- Demand paging
  - lazy swapper
  - strict demand paging
  - performance (EAT)
    - average page fault service
- Page fault handling
  - steps 1-6
- Page replacement algorithms
  - victim page
  - FIFO
  - LRU
    - counter/stack
    - approximation
    - (reference bit/second chance)
  - optimal
  - Belady's anomaly
- Frame allocation policy
  - fixed / priority
  - global/local
- Trashing
- Working-set
  - page-fault frequency scheme

NAMING

7 FILE SYSTEM (FS) ABSTRACTION

- Goal
  - physical characteristic
  - design implementation
- FS interface
  - metadata
- Naming
  - indirection
  - binding
  - model
    - name space
    - universe
    - mapping algorithm
    - ~~(IPv4)~~
  - operations
    - resolving
    - bind/unbind
    - enumeration / comparison
  - policy alternatives
- Types of lookup
  - table lookup
  - recursive lookup
    - pathnames
  - multiple lookups
    - search paths



- context
  - default/explicit
- Path names
  - forward / backwards
  - soft links
- Name discovery
- FS introduction
  - DAG (directed acyclic graph)
  - operations
  - no cycles
- Access control
  - row-wise: ACL
  - column-wise: Capabilities
  - POSIX
- Concurrency

8 FS IMPLEMENTATIONS

- file types
  - text/binary
- directories
  - linear list
  - hash table
  - B-tree
- open file interface
  - byte sequence
  - record sequence
  - key-based/tree structured
  - ~~memory-mapped~~
- On-disk data structure
  - disk addressing
  - overview
    - index structure / index granularity / free space management / locality heuristics
- FAT-32 (file allocation table)
- FFS (fast file system)
  - inode
  - bitmap for free space
  - block groups
    - first-fit allocation
- NTFS (new technology file system)
  - MFT (master file system)
  - metadata
    - table 0-11
- In-memory data structure
  - page cache
  - recovery
- Partitions
- Logical volumes
- VFS (virtual file system)

9 I/O SUBSYSTEM 1

- Recap device
  - bus/registers
  - interrupt
    - maskable
  - UART
  - DMA/PIO
  - I/O protection
  - IOMMU
- Device drivers
  - driver threads
    - 1. interrupt handler
    - 2. thread
    - 3. user
  - deferred procedure calls (DPC)
  - FLIH (1LIH)
  - SLIH (2LIH)
- I/O subsystem
  - caching / spooling / scheduling / buffering
  - naming
  - discovery
- Block devices
  - (USB, network, ...)
- Character devices (stream sensors)
  - (Screen, Keyboard, ...)
- Unix/Linux device configuration

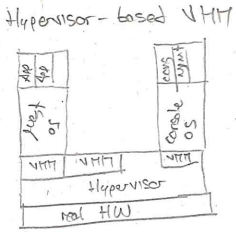
10 I/O SUBSYSTEM 2

- interface to network
  - protocol stack
- network stack implementation
  - receiving
  - sending
  - TCP (transmission control protocol)
  - triggers
  - actions
- Memory management
- Performance issues
  - TCP offload
  - buffering
  - interrupt coalescing / throttling
  - receive-side scaling
- NAPI (linux new API)
  - transmit path
  - receive path
  - transmit interrupt
    - ring empty

- ring occupancy drops
- buffer summary
- receive side scaling

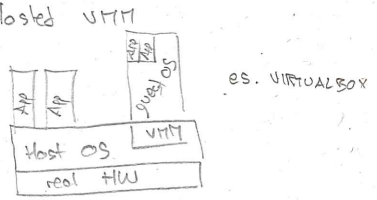
11 VIRTUAL MACHINE MONITORS (VMM)

- definition — virtualizes HW, illusion — Guest OS = GOS
- advantages
  - server consolidation
  - performance isol
  - backward comp
  - cloud computing
  - OS testing
  - something unter os
  - multiple OS
- Structure
  - Hosted VMM
  - Hypervisor-based VMM



- Virtualizing the CPU — strictly enabled over itself
  - x86 (not virtualizable)
  - solutions
    - emulation (slow, I/O)
    - paravirtualization — modify GOS
    - binary rewriting
    - hardware support

PUSHF / POPF (include interrupt) IF flag ⇒ different behaviour when privileged



- Virtualizing the MMU
  - Xen 3 supported modes
    - direct PT
    - shadow PT V→M (keep consist) → must check PT of GOS to avoid problems
    - HW assisted paging V→P, P→M
- Virtualizing the Memory
  - Allocating
  - Ballooning
    - taking/returning RAM
- Virtualizing Devices
  - paravirtualized devices (fake DD that talk to VMM)
  - ~~APIC~~ ?
- Virtualizing the Network
  - (4) options for where real drivers are
  - single-root I/O virtualization
  - PF, VF (physical/virtual function)
  - self-virtualizing device

PCI peripheral component interconnect

12 RELIABLE STORAGE, NUMA

- Properties
  - reliable → continues to store data
  - available → respond to req
- Transactions
  - RAID 1, 5
    - 3 ways of losing data
  - sector/page failure
    - error rate formula
  - device failure
    - MTTF (mean time to failure)
    - formula

RAID = data storage virtualization tech
- combine multiple physical Disks
- redundancy
- performance

① ⑤ keep pair, copy (redundancy)

$(1-\varepsilon)^{size}$

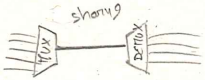
annual failure = 1/MTTF

- Ensure atomicity
- Recovery
  - terminology
  - MTTR (mean time to recover)
  - MTTDL (——— data loss)
- Hardware trends
  - comparison computer-human speed
- NUMA (non-uniform memory access)
  - striping hardware
- ~~How to compute fast~~
  - takeaways:
    - use bypass
    - reduce overhead

PART 2 – NETWORK

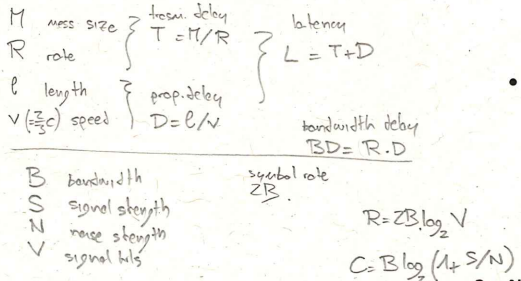
0 FUNDAMENTALS

- Multiplexing
  - standard, statistical
- Metcalfe's Law
- Scale
  - vicinity, building, city ~~man~~, country, planet
  - PAN, LAN, MAN, WAN, Internet
- Traceroute
- Protocols + Layering
  - vertical, horizontal, pros/cons
- OSI 7 Layer Reference model
  - physical, data link, network, transport, session, presentation, application
  - bit, frame, packet, segment, message
  - repeater, switch, router



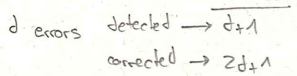
1 PHYSICAL LAYER

- Formulas
  - latency, rate, length, speed, message size
  - transmission delay, propagation delay, bandwidth delay
- Signal carriers
  - twisted pair wires, coaxial cables, fiber, wireless
- Effects
  - delay, attenuated, cutoff, noise – frequency, bandwidth
- Wireless multipath
- Modulation
  - bits, voltages, NRZ
  - clock recovery, data -> code
  - baseband, passband
- Limits
  - Nyquist, Shannon capacity
  - bandwidth, signal strength, noise strength, levels
  - bitrate, SNR, channel information carrying rate
  - DSL



2 DATA LINK LAYER

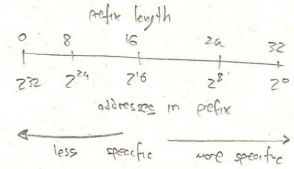
- Framing
  - counting bytes
  - byte stuffing, flag value, escape code, PPP, SONET
- Error Detection
  - reliability, redundancy
  - D data + R check bits
  - Parity bit, sum
  - Checksum (Internet) – send, receive
  - CRC (Cyclic Redundancy Check) – send, receive
- Error Correction
  - Hamming distance, code
  - syndrome, steps to compute



- (convolutional codes)
- (LDPC)
- Detection vs Correction
  - general pros/cons, random errors, burst errors
- Retransmission
  - ARQ, (ACK)
  - timeout, duplicates, Sliding Window
- Multiple Access MA
  - TDM, FDM (time/frequency division multiplexing)
  - randomized access, contention-free
  - (ALOHA network)
  - CSMA
  - CD
  - BEB
  - Classic Ethernet, frame format
  - Wireless
  - Hidden / Exposed terminals
  - MACA (multiple access collision avoidance)
  - RTS, CTS (request/clear to send)
  - Wi-Fi, AP (access point)
  - OFDM (orthogonal frequency division multiplexing)
  - CFMA, TTMAP
- Switching
  - ports wired, hubs
  - forwarding
  - loop avoidance
  - spanning tree MST, algorithm
  - shortcomings

3 NETWORK LAYER

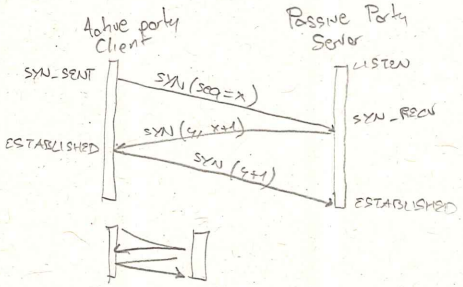
- Routing, Forwarding
- Network service models
- Datagrams (letters)
- Virtual Circuits (phone calls)
  - three phases
  - connection establishment, data transfer, teardown
  - MPLS (multi-protocol label switching)
- Pros/Cons
  - setup phase, router state, addresses, routing, failures, QOS (quality of service)
- Packets
- IP (internet protocol)
  - ISP (internet service provider)
  - IPv4 header
  - version, IHL, TOS, tot length, frag ID, DF MF, frag offset
  - TTL, protocol, checksum, source + destination address, options
  - prefixes notation, specificity, length L
  - public, private
  - allocation, IANA; RIR
  - forwarding, longest matching prefix



- other aspects (TTL, checksum, fragment, congestion, errors, options)
- DHCP (dynamic host configuration protocol)
- ARP
  - mapping MAC addresses
- MTU (max transmission unit), packet size
  - fragmentation, MF DF
  - path discovery
- ICMP
  - error report and testing
  - traceroute, TTL, PING
- IPv6
  - addresses format
- NAT (network address translation)
  - pros/cons
  - box, internal / external
  - firewall, intrusion detection
  - TCP (transmission control protocol)
  - downsides / upsides
  - FTP
- Routing Algorithms
  - Goals - correct, efficient, fair, fast, scalable
  - Rules – decentralized, alike, neighbors, concurrent
  - Best paths – latency, bandwidth, money, hops, topology
- Shortest Path Routing
  - cost function
  - sink tree
  - Dijkstra's algorithm
- Distance Vector routing
  - DV
  - RIP (routing info protocol)
  - Bellman-Ford algorithm
  - forwarding table computation
  - issues, split horizon, poison reverse
- Flooding
  - duplicate suppression
- Link-state routing
  - LSP (packets)
  - ageing
- ECMP (equal cost multi path) routing
- Hierarchical routing
  - regions
  - adjust size of prefixes
  - subnets
  - aggregation
- Multiple parties / Inter domain routing
  - networks, ISP, CDN, IXP
  - issues – scaling, policy
  - TRANSIT
  - PEER
  - BGP (border gateway protocol)
  - AS (autonomous system)

4 TRANSPORT LAYER

- aaa
- simo s2 sono innamoratissima di te
- TCP & UDP pro/cons
- TCP (transmission control protocol)
  - connections
  - MSS
  - 3 way handshake
  - ISN (initial sequence number)
  - state machine
  - connection release
  - TIME_WAIT
  - header



- UDP (datagrams)
  - voice IP, DNS, RPC DHCP
- Sockets
  - ports
  - bind, close
  - listen/accept/connect
  - sendto / receive
- Sliding window
  - stop-and-wait
  - go-back-N
  - selective-repeat
  - W
  - LAS, LAR, LFS
  - sequence number
  - SENDER/RECEIVER
    - recv()
  - pipelining
  - SACK (selective ack)
- Flow control
- Retransmission timers
- Congestion
  - bandwidth allocation (models)
  - challenges / solutions
  - max/min fairness
  - Avoidance
    - ECN (electronic communication network)
- AIMD (additive increase multiplicative decrease)
  - cwnd
  - ACK clocking
  - TCP slow start
    - problems / solutions
  - Fast retransmit
  - Fast recovery

- TCP reno

5 SESSION LAYER

6 PRESENTATION LAYER

7 APPLICATION LAYER

- DNS (domain name system)
  - names / addresses / resolution
  - protocol
- NIC
- hierarchical namespace
  - TLD (top level domains)
  - subdomains
  - ICANN (internet corp. for assigned numbers & names)
  - zone
  - SOA, A, AAAA, CNAME, MX, NS
  - resolve
    - recursive / iterative query
    - latency
- HTTP (hypertext transfer protocol)
  - steps
  - get / head / post / put / delete / trace / connect / options
  - codes 1XX-5XX
  - echo request
- PLT (page load time)
- CDN (content delivery network)
  - clever usage of DNS
- Proxy
  - ETag
  - web and caching
  - reusing local copy
- P2P / BitTorrent
  - challenge / solution
  - torrents
  - distribution tree
  - DHT (distributed hash tables)
  - tracker